Operator scaling and matrix sensing

Plan:
- Matrix & operator scaling recap.
- Regression, mirror descent, Matrix scaling
- Sensing, operator scaling, moment maps.
- Open directions

Overparametrized linear regression + matrix sensing.

- **Sparse regression:**

  $\mathcal{A}: \mathbb{R}^N \to \mathbb{R}^M$ linear map, $b \in \mathbb{R}^M$

- Want to solve $\mathcal{A}(A) = b$ for $A \in \mathbb{R}^N_{\geq 0}$.

- If if $N \gg M$, many solutions; want a sparse one.

- Proxy for sparsity: $\|A\|_1$.

- Minimize **quadratic loss** $\|\mathcal{A}(A) - b\|_2^2$,

- Use **nonconvex** reparametrization

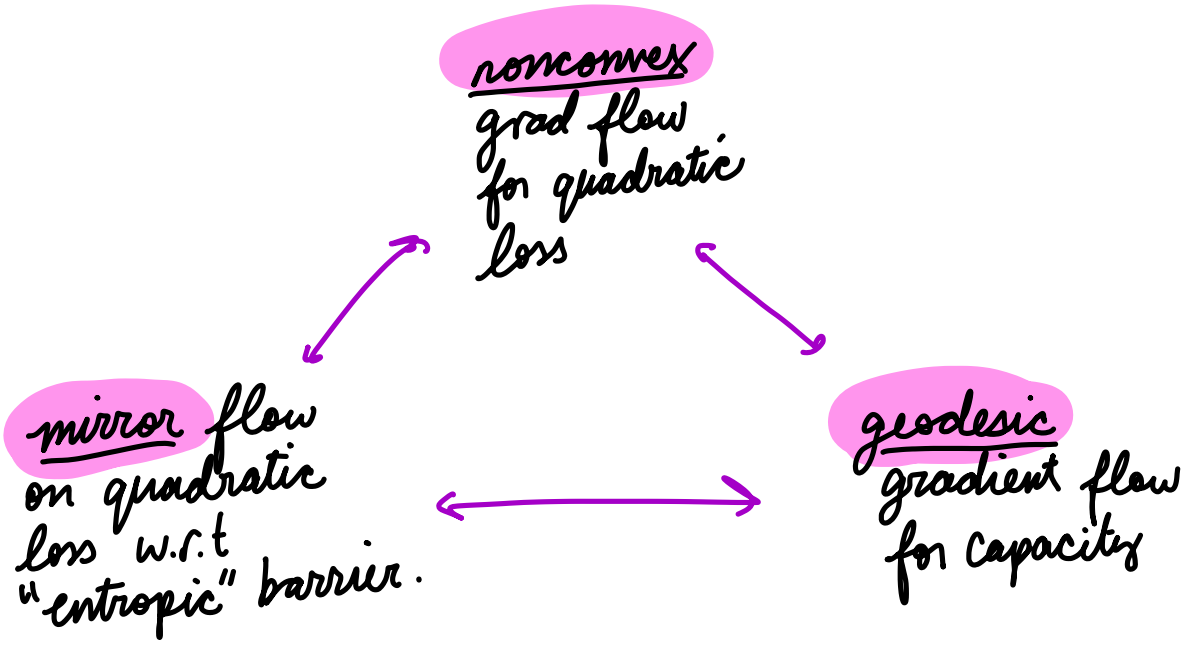
$$A = X \circ X$$

↑ coordinatewise product.

- **Matrix sensing**: similar thing,
  $\mathcal{A}: \mathrm{Mat}(N\times N) \to \mathbb{R}^M$
  want to solve $\mathcal{A}(\rho) = b$
  for $\rho \succeq 0$.
- In that case, reparametrize $\rho = XX^T$.
  for $X \in \mathrm{Mat}(N,k)$.

Three flows

- For regression case, identical:



- More detail to follow.
- **Main Question**: also true for matrix sensing?
- Interesting even for special case called **operator scaling**.
- **Example: Matrix scaling**

  Given $A \geq 0 \in \mathrm{Mat}(m,n)$,

  $(r,c) \in \mathbb{R}^m \times \mathbb{R}^n$,

  find diagonal matrices D_1, D_2 s.t.

  $\tilde{A} = D_1 A D_2$ has

  $$\tilde{A}\mathbb{1} = r, \quad \tilde{A}^T\mathbb{1} = c.$$

- **Rephrase**: Let

▷ $u(A) := (A\mathbb{1}, A^T\mathbb{1})$

▷ $b := (r, c)$

- So we are looking to solve $u(A) = b$ a very special way.

- Note $u : \mathbb{R}^{mn} \to \mathbb{R}^m \times \mathbb{R}^n$; $N = mn$, $M = m+n$

Mirror vs. nonconvex flow.

- Back to general $u : \mathbb{R}^N \to \mathbb{R}^M$.

- write $A = X \circ X$, let $X(t)$ gradient flow of $q(X) := \|u(X \circ X) - b\|_2^2$.

- To minimize $\|A\|_1$, Start at $X(0) = \alpha \mathbb{1}$ for $\alpha \to 0$.

Fact: Let $A(\infty) = \lim_{t \to \infty} X(t) \circ X(t)$; then

$$A(\infty) \xrightarrow[\alpha \to 0]{} \operatorname{argmin}_{\mu(A)=b} \|A\|_1 .$$

(Gunasekar et.al. 2017)

- Gunasekar et. al. conjectured same for matrix sensing; recently disproven! (Li, Luo, Lyu 2020)
- Example is 4x4; converges to lowest rank.

Stronger Fact: For any $A(0) \geq 0$,

$$A(\infty) = \operatorname*{argmin}_{u(A)=b} D_{KL}(A \,\|\, A(0))$$

* whenever min is finite.

where

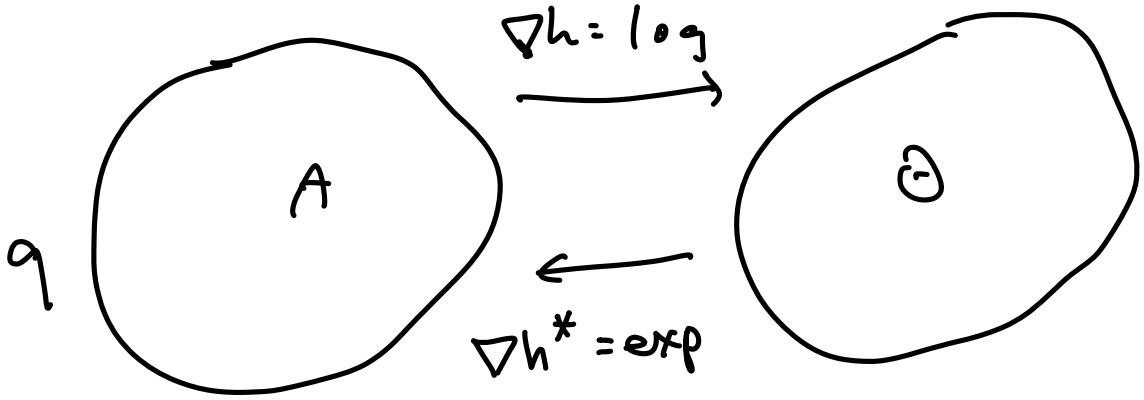
$$D_{KL}(A \,\|\, B) = \sum B_i - \sum A_i + \sum A_i \log \frac{A_i}{B_i}.$$

Proof Sketch: • suffices to show $A(t)$

mirror flow of loss q w.r.t. barrier

$h(A) = \sum A_i (\log A_i - 1)$ "negative entropy".

b/c D_{KL} is Bregman divergence w.r.t. h.



- flow: $\dot{\Theta} = -\nabla_A f(A)\big|_{A=\exp\Theta}$

$A(t) := \exp\Theta(t)$.

- both ODE's for A are

$$\dot{A}(t) = -\mu^*(\mu(A) - b) \circ A.$$

- Here $\mu^* : \mathbb{R}^M$ to $\mathbb{R}^N$ is adjoint of μ.

- Example: for matrix scaling,

$$\mu^*(y,z) = y\mathbb{1}^T + \mathbb{1}z^T. \qquad \square$$

Geodesic flow for matrix scaling

- capacity $\mathrm{cap}_{r,c}(A)$ defined by

$$\inf_{\substack{y\in\mathbb{R}^m \\ z\in\mathbb{R}^n}} \underbrace{\sum_{ij} A_{ij} e^{y_i+z_j} - y\cdot r - z\cdot c}_{:= f(y,z)}$$

- Sinkhorn is just alt. min;
scaling is $e^{\mathrm{diag}(y)} A\, e^{\mathrm{diag}(z)}$.

- Consider gradient flow for $f(x,y)$.

$$A(t) := e^{\operatorname{diag} y(t)} A\, e^{\operatorname{diag} z(t)}$$

- then $\dot{y}(t) = r - A(t)\mathbb{1}$

$$\dot{z}(t) = c - A(t)^T \mathbb{1}.$$

- Or with $b = (r,c)$, $\mu(A) = (A\mathbb{1}, A^T\mathbb{1})$,

$$(\dot{y},\dot{z}) = b - \mu(A).$$

- Now

$$\dot{A}(t) = \operatorname{diag}(\dot{y})A + A\operatorname{diag}(\dot{z})$$

$$= (\dot{y}\mathbb{1}^T + \mathbb{1}\dot{z}^T) \circ A$$

$$= \mu^*(\dot{y},\dot{z}) \circ A$$

$$= -\mu^*(\mu(A) - b) \circ A \,.$$

- Same as **mirror & nonconvex**!
- Nothing special about matrix scaling; generalizes for

$$f(Y) = \sum_i e^{\mu^*(Y)_i} A_i - Y \cdot b,$$

$$A(t) = e^{\mu^*(Y(t))} \circ A$$

- **Check** $e^{\mu^*(y,z)} \circ A = e^{\mathrm{diag}(y)} A e^{\mathrm{diag}(z)}$ for matrix scaling.

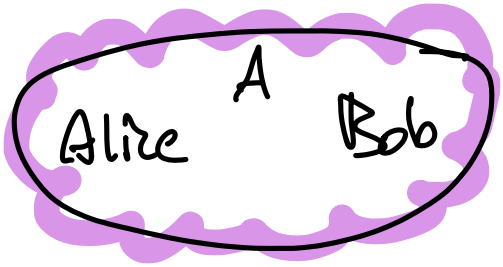
$\Rightarrow$ flow stays in group orbit!

Operator Scaling

Matrix Scaling

Nonnegative $n \times m$ matrix A

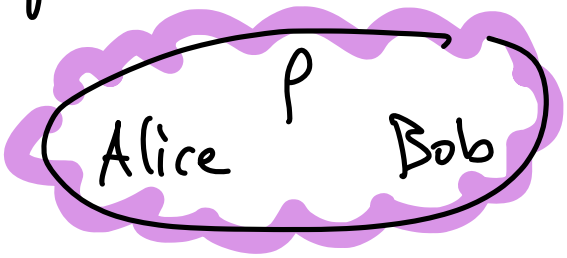
"(unnormalized) joint distribution" on $[n] \times [m]$



Operator Scaling

$mn \times mn$ PSD matrix ρ.

"unnormalized" joint quantum state on $\mathbb{C}^m \otimes \mathbb{C}^n$.



Correspondence: $A \longmapsto \rho := \sum_{ij} A_{ij} (e_i \otimes e_j)(e_i^* \otimes e_j^*)$

Marginals:

$$\mu(A) = (A^T \mathbb{1}, A\mathbb{1}) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^n_{\geq 0}$$

Alice ↓ Bob ↓

$$\mu(\rho) = (\mathrm{Tr}_2 \rho, \mathrm{Tr}_1 \rho).$$

$$\mu(\rho) \in \mathrm{PSD}(m) \times \mathrm{PSD}(n)$$

$$(\mathrm{Tr}_2 \rho)_{ij} = \mathrm{Tr}(E_{ij} \otimes I_n \rho)$$

$$(\mathrm{Tr}_1 \rho)_{ij} = \mathrm{Tr}(I_m \otimes E_{ij} \rho).$$

Scaling: $D_1 \succ 0, D_2 \succ 0$ diagonal $(D_1, D_2) \cdot A := D_1 A D_2$.	$g_1 \in GL(m)$ $g_2 \in GL(n)$ $(g_1, g_2)\, \rho$ $:= g_1 \otimes g_2 \, \rho \, g_1^\dagger \otimes g_2^\dagger$.

Problem (operator scaling)

Given $\rho \succeq 0$ $mn \times mn$,
target marg. $R \succeq 0$ $m \times m$ $C \succeq 0$ $n \times n$,
find scaling $\tilde{\rho}$ s.t.

$$\mathrm{Tr}_2 \tilde{\rho} = R \quad , \mathrm{Tr}_1 \tilde{\rho} = C.$$

- **Why?** Null-Cone, MLE's, quantum state estimation (quantum or ??).

Nonconvex v.s. geodesic for operator scaling.

- Nonconvex: write $P = XX^\tau$

  Now $q(X) = \| \mu(X) - b \|_F^2$.

- In general for matrix sensing,

$$\dot{X} = -\mu^*(\mu(X) - b)X.$$

ignoring constants.

- For operator scaling, $N = n$, $M = m^2 + n^2$.

$$\mu : \mathrm{Mat}(mn, mn) \to \mathrm{Mat}(m) \times \mathrm{Mat}(n)$$

$$\mu^*(Y, Z) = Y \otimes I_n + I_m \otimes Z.$$

- u^* is a **Lie algebra action** of $\mathfrak{gl}_n \times \mathfrak{gl}_m$ on $\mathrm{Mat}(nm, K)$.
- In particular, u^* is differential of **Lie group action**

$$(g, h) \cdot X = (g \otimes h) X .$$

of $GL_n \times GL_m$ on $\mathrm{Mat}(nm, K)$.

i.e. $e^{u^*(Y,Z)} = e^Y \otimes e^Z$.

- **Corollary**: For u of operator scaling,

$$x(t) = g(t) \times h(t) X \quad \forall t \geq 0 ,$$

hence $\rho(t) = x(t) x(t)^\dagger$ is a

Scaling of $\rho = \rho(0)$.

Geodesic flow:

- Capacity simpler for $R = C = I_n$ case, i.e. $b = (I_n, I_n)$.

$$\mathrm{Cap}(\rho) := \inf_{\substack{Y \succ 0 \\ Z \succ 0}} \underbrace{\mathrm{Tr}\, Y \otimes Z \rho - \log\det Y - \log\det Z}_{f(Y,Z)}.$$

- Write $Y = g^t g$, $Z = h^t h$; geodesics then Y are $Y(t) = g^t e^{Ht} g$ H hermitian.

- Riemannian gradient of $f(Y,Z)$ is then

$$\langle \nabla f, (H_1, H_2) \rangle_F = \partial_{t=0} f(g^t e^{H_1 t} g, h^t e^{H_2 t} h)$$

end up with $\boxed{\begin{aligned} \nabla_1 f &= \mu_1(\tilde{\rho}) - I \\ \nabla_2 f &= \mu_2(\tilde{\rho}) - I \end{aligned}}$.

for $\tilde{\rho} = g \otimes h \, \rho \, g^+ \otimes h^+$.

- Define

$$\dot{g}(t) = -\frac{\nabla f_1(Y,Z)}{2} g(t)$$

$$\dot{h}(t) = -\frac{\nabla f_2(Y,Z)}{2} h(t) \,,$$

(need this def because g not det. by Y).

$\Rightarrow$ $X(t) = g(t) \otimes h(t) \, X$,

has $\dot{X} = -\frac{1}{2} \mu^*(\mu(XX^T) - b) \, X$.

- **Moral**: same flow from geodesic capacity min & noncommex matrix scaling flow.
- First need $\mathfrak{u}^*$ to be Lie algebra representation!
- Doesn't converge to lowest rank solution - if $\rho(0) = A \otimes B$,

  $\rho(t) = A(t) \otimes B(t). \ \forall t.$

  $\mu(\rho) = (A \operatorname{Tr} B, B \operatorname{Tr} A)$

  $\Rightarrow \rho(t) \to I \otimes I$; full rank n^2

  but always $\exists$ rank 1 solution.
- For 3-tensor scaling w/ $\operatorname{rk} X = 1$ it can take very long to converge

Question: can we also interpret as a mirror flow?

Should tell us where it converges in terms of some Bregman div.

- What about matrix sensing NOT coming from lie groups?

$$f(Z) = \mathrm{Tr}\, e^{\mathcal{A}^*(Z)} \rho - b \cdot Z \; ?$$

Problem: e^Z makes no sense if Z not matrix or Lie gp. elt.

- What about $R \neq I$, $C \neq I$?

Other R, C

- gets more complicated.

- Want something like

$$\mathrm{cap}_{R,C}(\rho) = \inf_{Y,Z} \mathrm{Tr}\, Y \otimes Z\, \rho - \mathrm{Tr}\, R \log Y - \mathrm{Tr}\, C \cdot \log Z,$$

but the last terms are bad.

- we'd like the geodesic gradients to satisfy

eg. $\nabla_Y \mathrm{Tr}\, R \log Y = R$;

unfortunately <u>not</u> the case.

- But there is a nice g-concave function

$Y \mapsto \log\det(Y, R)$ which satisfies this.

- The catch: requires specific choice of coords. for tangent space, namely $Y = g^\dagger g$, $Z = h^\dagger h$, for g, h <u>lower tri</u>.

- Always exists; Cholesky decomp!

- Set

$$f(Y,Z) = \mathrm{Tr}\, Y \otimes Z\, \rho - \log\det(Y,R) - \log\det(Z,C)$$

- Now geodesic grad (taken this way) gives

$$\mu(\tilde{\rho}) - (R,C),$$

which is exactly what we need...

- Why not e.g.

$$\dot{g} = \frac{1}{2}\left(\mu_1(\tilde{\rho}) - R\right) g$$

like before? Not <u>lower tri</u>!

- must project it to lower tri ....

$$\dot{q} = \Pi_{\frac{1}{2}}\left(\mu_1(\tilde{\rho}) - R\right)q$$

- Thus $X(t)$ from nonconvex & $g(t) \otimes h(t) X$ from geodesic slightly different.

- Experimentally they converge to same ρ.

- Any fix? Interestingly even for full observation ($m = 1, n > 1$).